NCBI Pathogen Detection

From next generation sequencing reads to clusters of related bacterial pathogens based on SNPs

https://www.ncbi.nlm.nih.gov/pathogens/

National Center for Biotechnology Information • National Library of Medicine • National Institutes of Health • Department of Health and Human Services

Scope

Bacterial pathogens that originate from a food source are a serious concern in the US, with estimates by CDC that each year roughly 1 in 6 Americans (or 48 million people) get sick, 128,000 are hospitalized, and 3,000 die of foodborne diseases. Outbreaks are defined where two or more people become ill when ingesting the same food. If the bacterial pathogens that cause the illness are genetically related, it becomes easier to trace the source of the illness when samples from both patients and food source are compared. US Federal agencies involved in public health for foodborne illnesses have agreed to move forward on this by using whole genome sequencing. A pilot project was started in 2013 for *Listeria* with all federal agencies responsible for food safety (CDC, FDA, and USDA) sequencing all *Listeria* isolates, whether in clinical patients or identified during inspection of food products or processing facilities, and depositing the data in the public archives at NCBI. This project demonstrated that whole genome sequencing could be used to increase the number of clusters detected, with a decrease in median cluster size, and with more outbreaks solved [1]. The goal is to extend this to all 90,000 foodborne bacterial pathogens (*Campylobacter*, *Listeria*, *Escherichia coli* and *Shigella* (STECs), and *Salmonella*) that are collected in the US every year and sequenced in real time by the end of 2018.

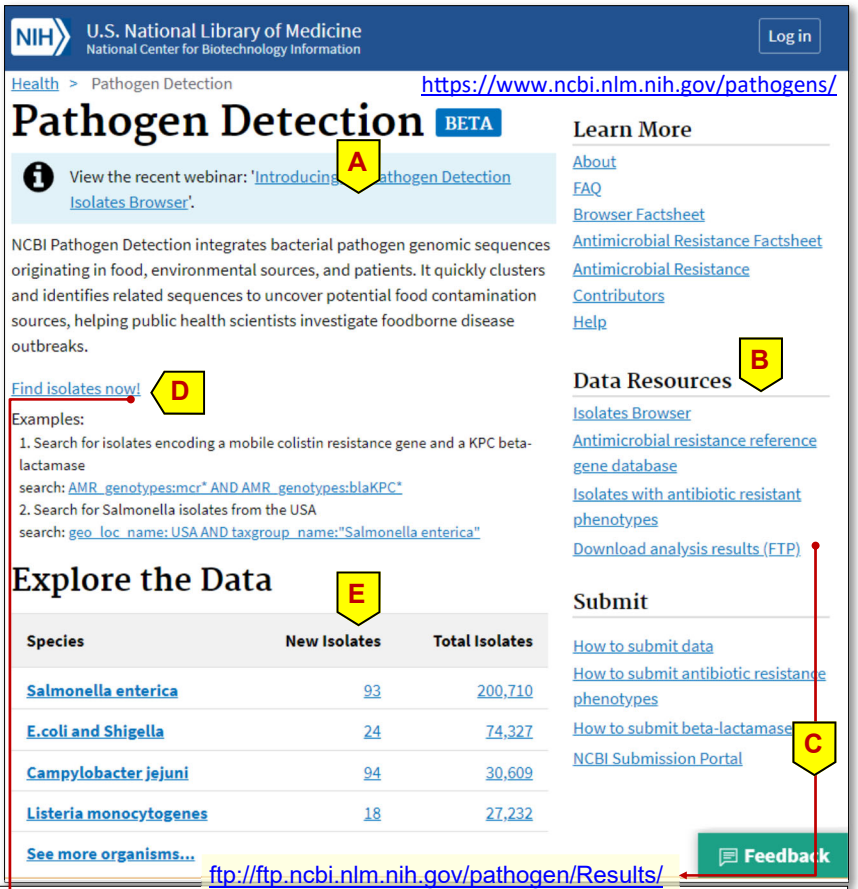
The NCBI Pathogen Detection pipeline takes the incoming sequencing data and assembles, annotates, and clusters the genomes together to facilitate the analysis of genetically related strains to aid outbreak and traceback investigations. Phylogenetic trees are constructed for each SNP cluster using maximum compatibility [2]. Currently clusters are calculated daily for each organism group if new data are submitted and released publicly. The scope of the project has now expanded to include non-foodborne pathogens, especially those that are becoming increasingly antimicrobial resistant (AMR). The workflow uses a reference set of acquired resistance genes/proteins to report for each isolate the AMR genes that are encoded in the genome sequence. This database of antimicrobial resistance pathogens is part of the National Action Plan for Combating Antimicrobial-Resistant Bacteria by providing a resource for the research community on resistant organisms that have genomic sequence data [3].

Data Access

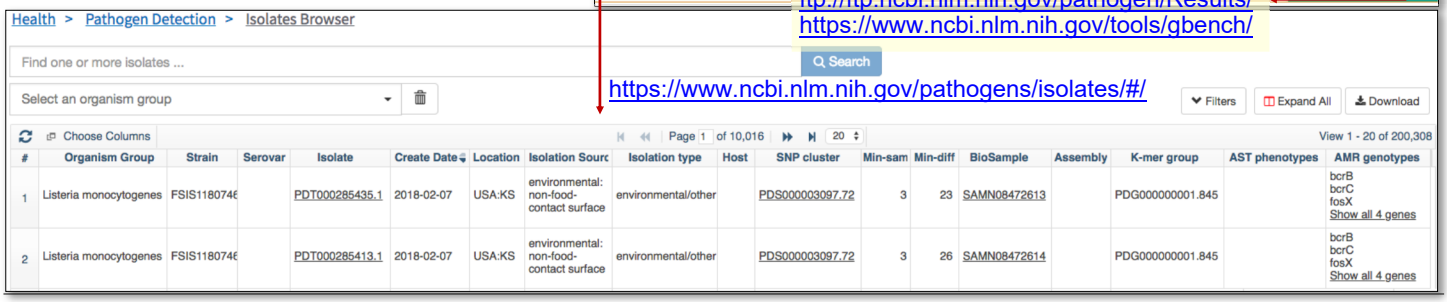
Several venues are available for accessing data from Pathogen Detection Pipeline:

1. The Pathogen Detection homepage (**A**) provides links both to the Isolate Browser and FTP results (**B**).
2. The FTP site (**C**) offers access to all of the phylogenetic trees and metadata tables. You can use NCBI's Genome Workbench tool to open and explore the files locally. See the readme for more information about the FTP.
3. The Isolate Browser (**D**) allows search and browse. In its display, each row is an assembled genome (either from SRA data by the Pathogen Detection Pipeline or genomes deposited in GenBank).

The table (**E**) on the main page displays total clusters by organism group and isolates for those groups that were newly deposited since the last cluster calculation.



ftp://ftp.ncbi.nlm.nih.gov/pathogen/Results/
https://www.ncbi.nlm.nih.gov/tools/gbench/



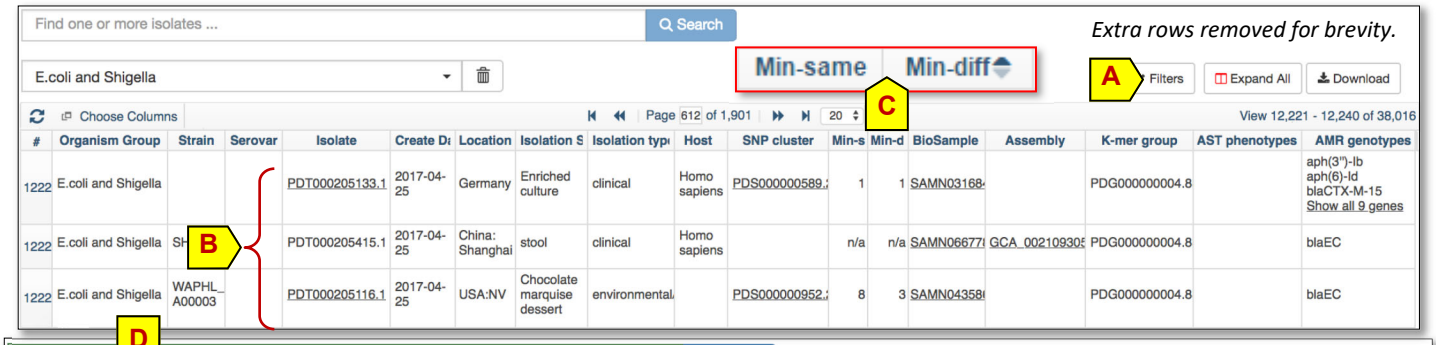
https://www.ncbi.nlm.nih.gov/pathogens/isolates/#/

Contact: pd-help@ncbi.nlm.nih.gov

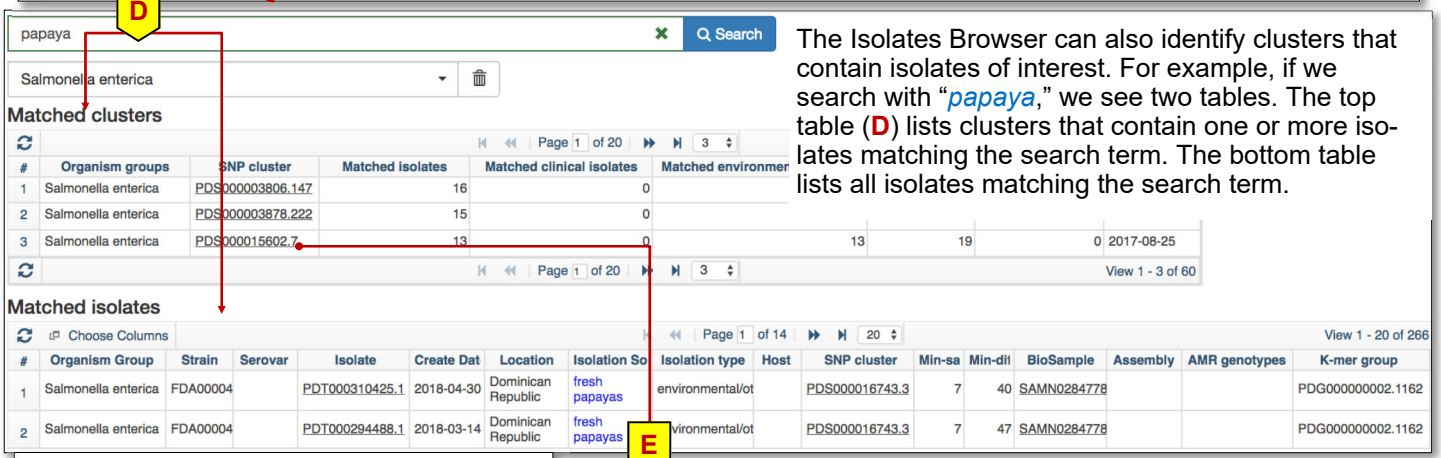
Usage Examples

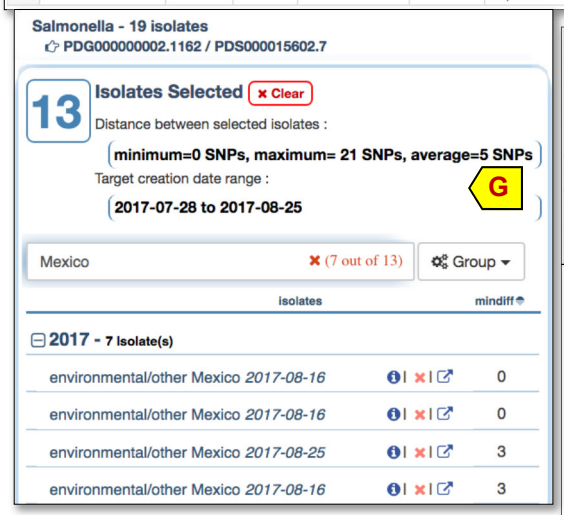
1) Isolate Browser: Navigation and searching for cluster of interest

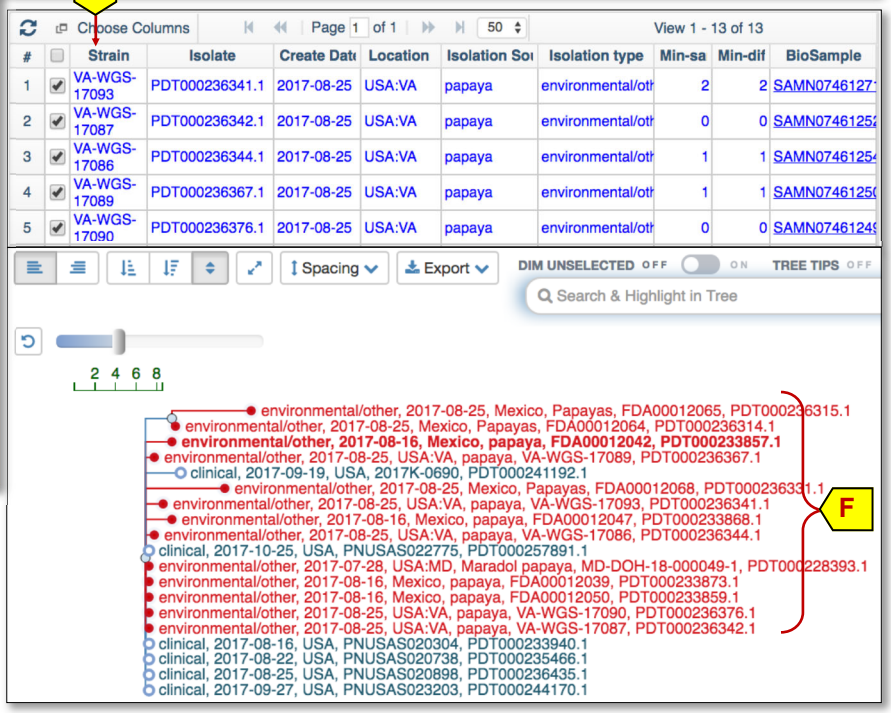
From the Pathogen Detection homepage's table (p1, **E**), you can click on the number in the 'New Isolates' column for an organism group to locate new clusters for that group. You can search directly in the Isolate Browser using structured terms. For *E. coli* and *Shigella* group, the structured query terms are: *taxgroup_name:"E.coli and Shigella" AND new:1* (http://bit.ly/ncbi-pathogen-1). The search results will change every time new data is deposited. You can toggle open the Filters section (**A**, more on this in p 3) to narrow down isolates displayed. In the Isolate Browser, the **Min-same** and **Min-diff** columns (insert) report on the SNP differences for each isolate (**B**) based on Isolation type (clinical vs. environmental). Sorting on the **Min-diff** column (**C**) in ascending order brings isolates with the shortest SNP distance between a clinical and environmental samples to the top. This similarity is critically important for public health labs.



The Isolates Browser can also identify clusters that contain isolates of interest. For example, if we search with "*papaya*," we see two tables. The top table (**D**) lists clusters that contain one or more isolates matching the search term. The bottom table lists all isolates matching the search term.





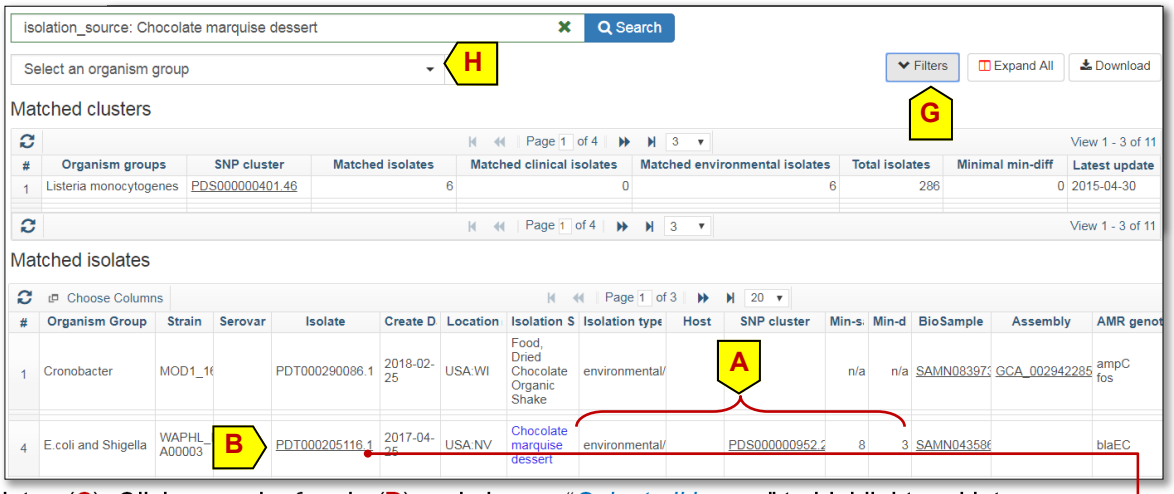


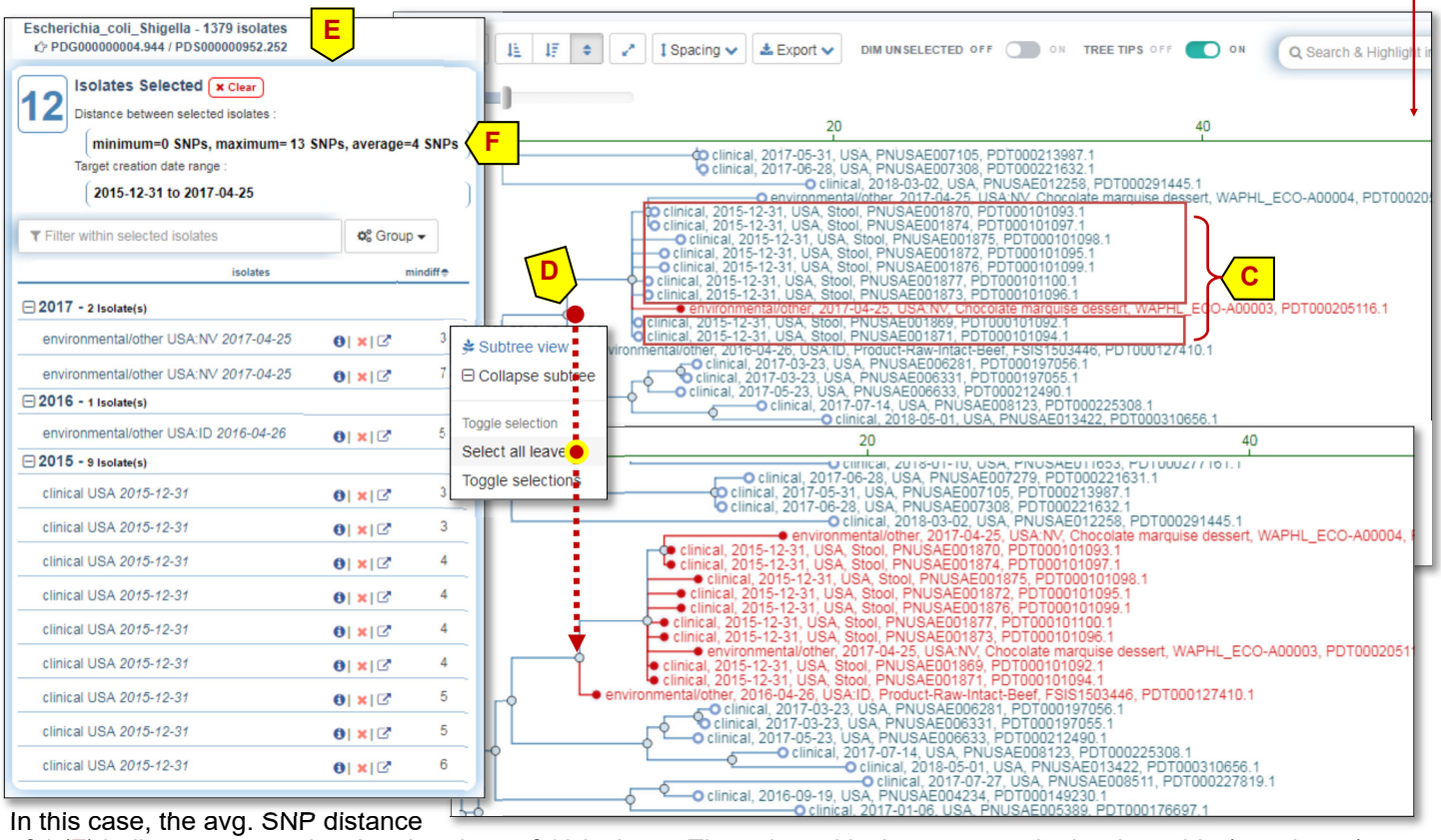
Selecting one cluster, PDS000015602.7 in this case, we arrive on a page where the table lists those isolates within this cluster (**E**) that match the term "*papaya*" Below that is a tree with all of the isolates hitting the search term highlighted in red (**F**). The control column to the left (**G**) reports the minimum and maximum SNP distances. We also can filter the isolates selected by custom input, such as "*Mexico*" to limit the list of isolates

Usage Examples

2) Use Case One: To identify potential clusters of interest for public health

We will use an environmental isolate PDT000205116.1 in this exercise, which is 3 SNPs away from a clinical isolate (**A**). Click the isolate identifier (**B**) to open the SNP Tree Viewer, which zooms to the particular branch of the phylogenetic tree where that isolate is located. Our environmental isolate (automatically highlighted) is clustered with several clinical isolates (**C**). Click a non-leaf node (**D**) and choose "*Select all leaves*" to highlight and interrogate that branch in the SNP distance tree. The panel in the left column (**E**) updates upon such selection to provide a summary of the selected branch along with a timeline of isolation.





In this case, the avg. SNP distance of 4 (**F**) indicates a very closely related set of 12 isolates. The selected isolates appear in the datatable (not shown) above the tree to display the metadata associated with these isolates. The clinical isolates were from 2015, while the environmental isolates were from raw beef in 2016 and desserts in 2017. A public health professional reviewing this data needs to make a determination based on additional metadata if further investigation is warranted.

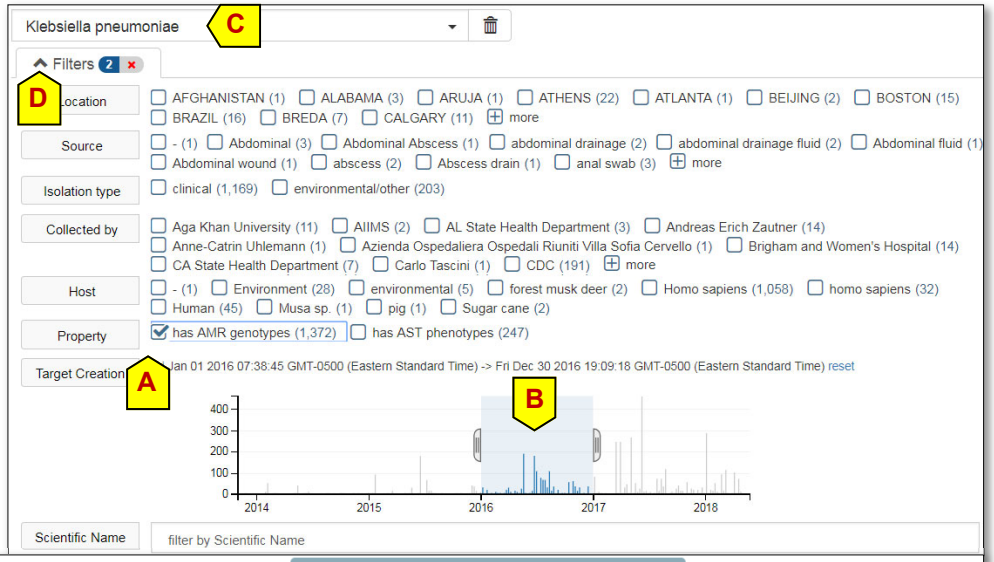
3) Use Case Two: To identify isolates encoding antimicrobial resistance genes

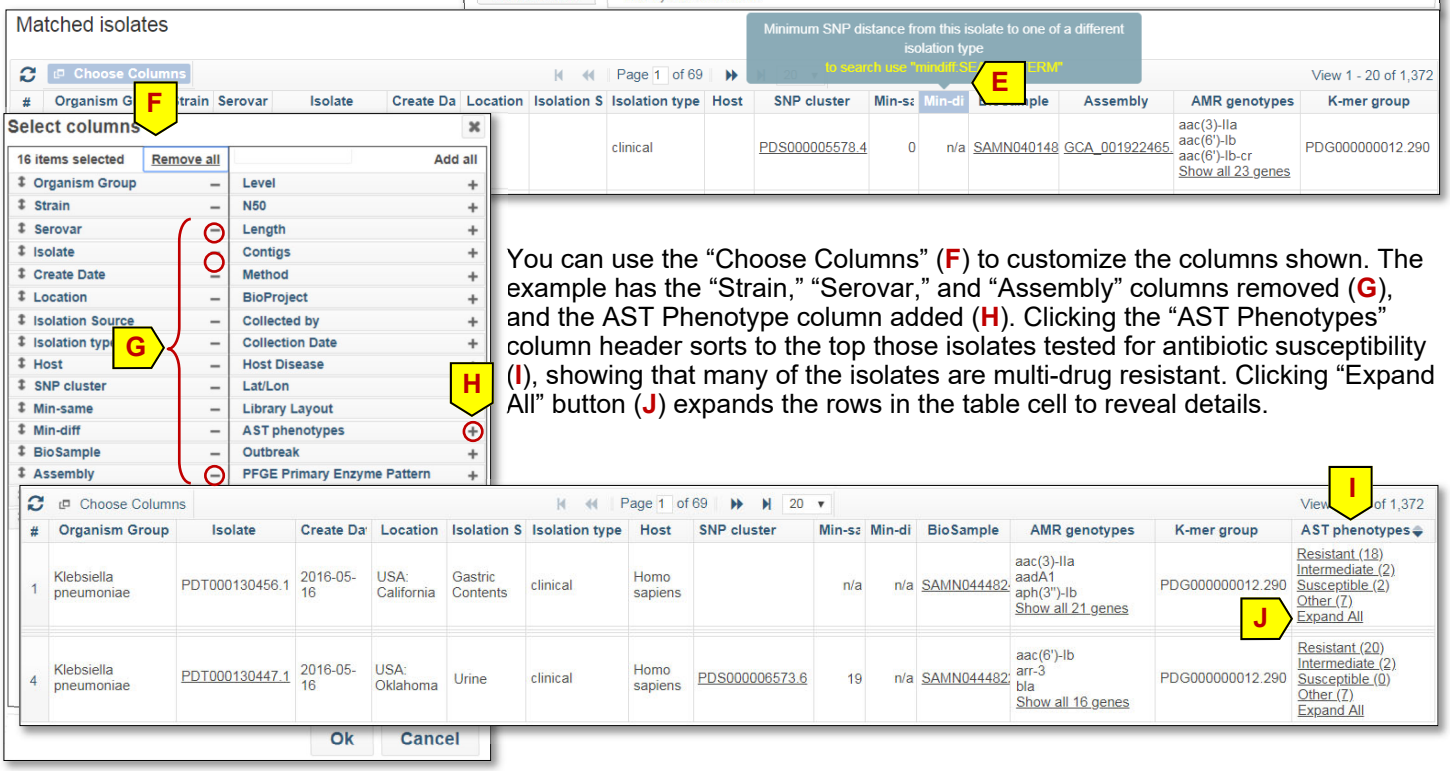
The Filter button (**G**) above the cluster table provides additional options to allow filtering isolated by their isolation location and source, sample collecting lab, host, whether the isolate has antimicrobial resistance genes or has antibiotic susceptibility test phenotypic data, a date range control, and by the scientific name of the isolate. In this example, we focus on a subset of isolates by selecting "Klebsiella pneumoniae" from the "Select an organism group" pull-down menu (**H**) to the left of the Filter button. Click the Filter button to activate the display and access preset filter options to further refine the isolate list.

3) Use Case Two (cont.)

In the Filter control (right) check "*has AMR genotypes*" (**A**), and set date range slider to 2016 (**B**). Together with "*Klebsiella pneumoniae*" organism group selection (**C**), the setting retrieves K. pneumoniae isolates released in 2016 that encode an antimicrobial resistance gene. Click the up arrow (**D**) to close the Filter control.

Matched isolates table updates upon filter activation. Hover over a column header to see the full column description and search syntax (**E**).





You can use the "Choose Columns" (**F**) to customize the columns shown. The example has the "Strain," "Serovar," and "Assembly" columns removed (**G**), and the AST Phenotype column added (**H**). Clicking the "AST Phenotypes" column header sorts to the top those isolates tested for antibiotic susceptibility (**I**), showing that many of the isolates are multi-drug resistant. Clicking "Expand All" button (**J**) expands the rows in the table cell to reveal details.

Future Work

The NCBI Pathogen Detection team is working on a number of enhancements to the interface including:
1) genomic locations of resistance genes, 2) ability to search using specific MIC values, and 3) reporting of biocide and metal resistance genes, and virulence genes.

References

1. Allard et al. (2016). The Practical value of Food Pathogen Traceability through BUILDING a Whole-Genome Sequencing Network and database. J Clin Microbiol. 2016 Mar 23. pii: JCM.00081-16.
2. Cherry J. (2017). A practical exact maximum compatibility algorithm for reconstruction of recent evolutionary history. BMC Bioinformatics. 2017 Feb 23;18(1):127. doi: 10.1186/s12859-017-1520-4.
3. Jackson et al. (2016). Implementation of Nationwide Real-time Whole-genome Sequencing to Enhance Listeriosis Outbreak Detection and Investigation. Clin Infect Dis. 2016 Apr 18. pii: ciw242.
4. National Action Plan for Combating Antimicrobial-Resistant Bacteria.
http://go.usa.gov/xQpax and http://go.usa.gov/xQpaa
5. If you wish to submit data to the NCBI Pathogen Detection pipeline, please review the submission page:
https://www.ncbi.nlm.nih.gov/pathogens/submit/